The Role of Context in Affective Behavior Understanding

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292
morency@ict.usc.edu

1 Introduction

Face-to-face communication is highly interactive. Even when only one person speaks at the time, other participants exchange information continuously amongst themselves and with the speaker through gesture, gaze, posture and facial expressions. Such affective feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participants ability to communicate [2, 20]. During multi-party interactions such as in meetings, information is exchanged between participants using both audio and visual channels. Visual feedback can range from a simple eye glance to a large arm gesture or posture change. One important visual cue is head nod during conversation. Head nods are used for displaying agreement, grounding information or during turn-taking [7, 8]. Recognizing these affective gestures is important for understanding all the information exchanged during a meeting or conversation, and can be particularly crucial for identifying more subtle factors such as the effectiveness of communication [17], points of confusion, status relationships between participants [18], or the diagnosis social disorders [15].

This chapter argues that it is possible to significantly improve state-of-the art recognition techniques by exploiting regularities in how people communicate. People do not provide affective feedback at random. Rather they react to the current topic, previous utterances and the speaker's current verbal and nonverbal behavior [1]. For example, listeners are far more likely to nod or shake if the speaker has just asked them a question, and incorporating such dialogue context can improve recognition performance during human-robot interaction [11]. More generally, speakers and listeners co-produce a range of lexical, prosodic, and nonverbal patterns. Our goal is to automatically discover these patterns using only easily observable features of human face-to-face interaction (e.g. prosodic features and eye gaze), and exploit them to improve recognition accuracy.

This chapter shows that the recognition of affective gestures can be im-

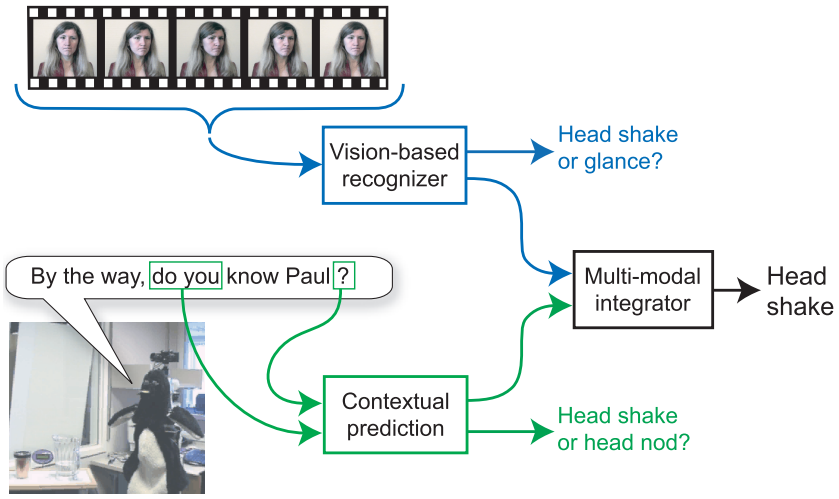

Figure 1: Contextual recognition of head gestures during face-to-face interaction with a conversational robot. In this scenario, contextual information from the robot's spoken utterance helps disambiguating the listener's visual gesture.

proved by considering the behaviors of other participants in the conversation. Specifically, it shows that the multimodal context from the current speaker can improve the visual recognition of listener gestures. We introduce the idea of encoding dictionary, a technique for contextual feature representation inspired by the influence speaker context has on the listener feedback. Automatic selection of relevant contextual features is performed by looking at individual and joint influences of context. The final contextual integration is done using a discriminative sequential model. We show the importance of context in affective behavior understanding on two different domains: interaction with a robot and human dyadic interaction. An important contribution of this chapter is the introduction of co-occurrence graphs which models co-occurrence between contextual cues such as spoken words and pauses, and affective gestures. By analyzing these co-occurrence patterns, we show how to automatically select relevant contextual features and predict when visual gestures are most likely. Our context-based recognition framework allows us to predict, for example, that in certain contexts a glance is not likely whereas a head shake or nod is (as in Figure 1), or that a head nod is not likely and a head nod misperceived by the vision system can be ignored.

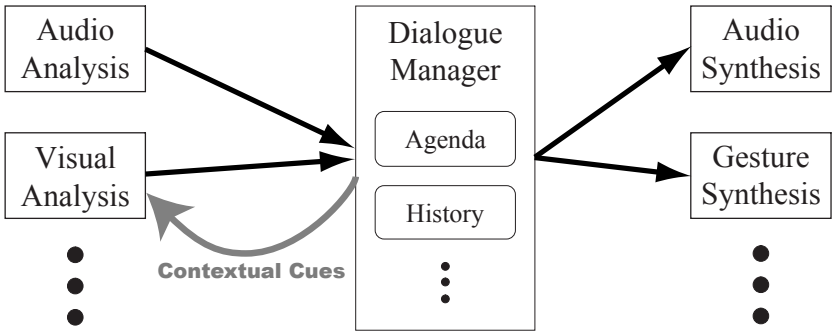

Figure 2: Simplified architecture for embodied conversational agent. Our method integrates contextual information from the dialogue manager inside the visual analysis module.

2 Context Definition

In this section we present how context can be defined during interactions with an embodied conversational agent such as a robot or virtual human, or during interactions between two or more human participants.

2.1 Robot and Agent Interactions

Figure 2 is a general view of the architecture for an embodied conversational agent[1]. In this architecture, the dialogue manager contains two main subcomponents, an agenda and a history.The agenda keeps a list of all the possible actions the agent and the user (i.e. human participant) can do next. This list is updated by the dialogue manager based on its discourse model (prior knowledge) and on the history. Some useful contextual cues can be estimated from the agenda:

- What will be the next spoken sentence of our embodied agent?
- Are we expecting some specific answers from the user?
- Is the user expected to look at some common space?

The history keeps a log of all the previous events that happened during the conversation. This information can be used to learn some interesting contextual cues:

- How did the user answer previous questions (speech or gesture)?
- Does the user seem to understand the last explanation?

[1]In our work we use the COLLAGEN conversation manager [14], but other dialogue managers provide these components as well.

Based on the history, we can build a prior model about the type of visual feedback shown by the user. Based on the agenda, we can predict the type of visual feedback that will be shown by the user.

The simplified architecture depicted in Figure 2 highlights the fact that the dialogue manager already processes contextual information in order to produce output for the speech and gesture synthesizer. The main idea is to use this existing information to predict when visual feedback gestures from the user are likely. Since the dialogue manager is already merging information from the input devices with the history and the discourse model, the output of the dialogue manager will contain useful contextual information.

We highlight four types of contextual features easily available in the dialogue manager:

Lexical features Lexical features are computed from the words said by the embodied agent. By analyzing the word content of the current or next utterance, one should be able to anticipate certain visual feedback. For example, if the current spoken utterance started with "Do you", the interlocutor will most likely answer using affirmation or negation. In this case, it is also likely to see visual feedback like a head nod or a head shake. On the other hand, if the current spoken utterance started with "What", then it's unlikely to see the listener head shake or head nod–other visual feedback gestures (e.g., pointing) are more likely in this case.

Punctuation We use punctuation features output by the dialogue system as a proxy for prosody cues. Punctuation features modify how the text-to-speech engine will pronounce an utterance. Punctuation features can be seen as a substitute for more complex prosodic processing that are not yet available from most speech synthesizers. A comma in the middle of a sentence will produce a short pause, which will most likely trigger some feedback from the listener. A question mark at the end of the sentence represents a question that should be answered by the listener. When merged with lexical features, the punctuation features can help recognize situations (e.g., yes/no questions) where the listener will most likely use head gestures to answer.

Timing Timing is an important part of spoken language and information about when a specific word is spoken or when a sentence ends is critical. This information can aid the ECA to anticipate visual grounding feedback. People naturally give visual feedback (e.g., head nods) during pauses of the speaker as well as just before the pause occurs. In natural language processing (NLP), lexical and syntactic features are predominant but for face-to-face interaction with an ECA, timing is also an important feature.

Gesture display Gesture synthesis is a key capability of ECAs and it can also be leveraged as a context cue for gesture interpretation. As described in [4], visual feedback synthesis can improve the engagement of the user with the ECA. The gestures expressed by the ECA influence the type of visual feedback

from the human participant. For example, if the agent makes a deictic gesture, the user is more likely to look at the location that the ECA is pointing to.

2.2 Human Interactions

Communication is a joint activity and social scientists have long argued that it cannot be properly recognized and understood by focusing on participants in isolation but rather one must see individual behaviors within the context of the group or dyad [3, 6]. Translating this proscription to the domain of gesture recognition, this argues that features outside of the person-of-interest should correlate with their behavior, and representing and exploiting these contextual features should improve recognition accuracy. Here, we explore this idea within the domain of dyadic and multi-party conversations, specifically we consider whether adding contextual information about a speaker's behavior improves the ability to detect feedback gestures produced by listeners.

Our goal is to quantify the relationship between contextual information and visual gestures by looking at the time distribution of visual gestures given a contextual event. In our case, a contextual event can be a spoken word, a pause or the end of a sentence. If a relationship exists between a contextual event and a specific visual gesture (e.g., head nod) then we will expect to see a structure in the relative distribution. If no relationship exist, the relative distribution should be random.

As our interest is in producing online (real-time) recognition systems, we focus on contextual features that would be readily available to a real-time system (i.e, surface behaviors rather than the privileged mental state of individual participants). We define context as the set of events happening from other sources than the person of interest. For example, in a multi-party conversation between four people, we define context for one participant as the set of events coming from the three other participants. Since our goal in this chapter is to recognize affective gestures, we focus on context events related to spoken utterances:

Prosody Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [12]. For example, Ward and Tsukahara [19] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. We encode the following prosodic features, including standard linguistic annotations and the prosodic features suggested by Ward and Tsukhara [19]:

- Downslopes in pitch continuing for at least 40ms
- Regions of pitch lower than the 26th percentile continuing for at least 110ms (i.e., lowness)
- Utterances longer than 700ms
- Drop or rise in energy of speech (i.e., energy edge)

- Fast drop or rise in energy of speech (i.e., energy fast edge)
- Vowel volume (i.e., vowels are usually spoken softer)
- Lengthened words (e.g., "I li::ke it")
- Emphasized or slowly uttered words (e.g., "ex_a_c_tly")
- Words spoken with continuing intonation
- Words spoken with falling intonation (e.g., end of an utterance)
- Words spoken with rising intonation (i.e., question mark)

Pauses Listener feedback often follows speaker pauses or filled pauses such as "um" (see [5]). To capture these possible associations, we use the following contextual features:

- Pause in speech (i.e., no speech)
- Filled pause (e.g. "um")

Gaze and Gestures Gestures performed by the speaker are often correlated with listener feedback [3]. Eye gaze, in particular, has often been implicated as eliciting listener feedback. Thus, we encode the following contextual feature:

- Speaker looking at the listener

3 Context and Affective Behaviors

To study the relationship between context and affective gestures we look at the annotations from the AMI meeting corpus [**?**]. This corpus contains 46 meetings with annotated head gestures and spoken words of all four participants[2]. Each meeting varies between 20-40 minutes. The corpus contains follow-up meetings with the same participants. These series usually contain 3 or 4 meetings.

Participants were video recorded using a frontal camera and a close-talking microphone. The video sequences were manually annotated with spoken words, punctuation and head gestures (head nods and head shakes). The dataset contains 9745 head nods and 1279 head shakes. In our analysis, we used a total of 184 sequences (some meetings had only 3 participants annotated with head gestures).

[2]The corpus contains a larger number of meetings but we used only the meetings that had both head gestures and spoken words annotated

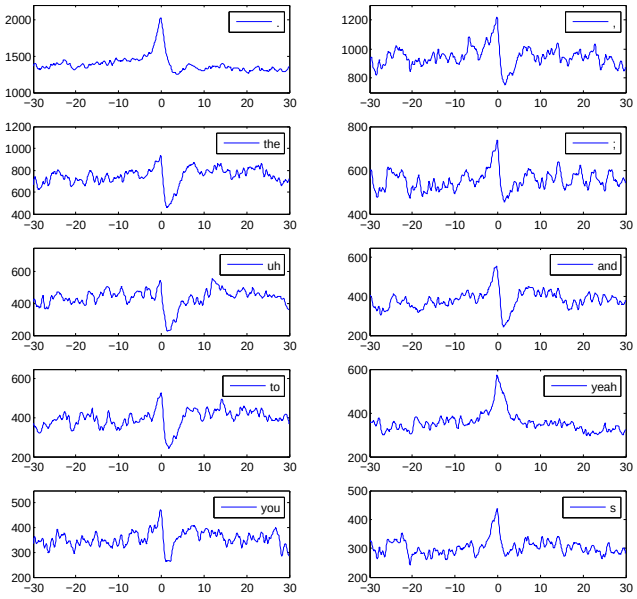

Figure 3: Examples of co-occurrence graphs. When analyzing this relationship between head nods and contextual events, three temporal patterns appear: ignition, transition and negation.

3.1 Co-Occurrence Graphs

Our goal is to analyze the relationship between contextual events and affective gestures. Our approach is to create a co-occurrence graph for each contextual event and each possible type of affective gesture. The co-occurrence graph, centered at the contextual event, represents how many visual gesture instances happened around that event. The co-occurrence graphs can be seen as temporal generalization of the co-occurrence matrices introduced by Haralick *et al.* [9].

For each instance of a contextual event, we slide a window of 0.1 second from -30 second before the event to 30 seconds after the event. If a visual gesture happens during a specific time window, the corresponding bin in the co-occurrence graph is incremented. By doing this for each instance of a specific contextual event, we get a time distribution of visual gesture given the contextual event. Figures 3 shows examples of co-occurrence graphs for different contextual events.

Figures 4 show cumulative co-occurrence graphs for head nods . The cumulative co-occurrence graph for head nods shows an interesting point: most of the relationship between head nods and contextual event seems to happen between -5 and 5 seconds. Past this time, the relationship seems mostly random.

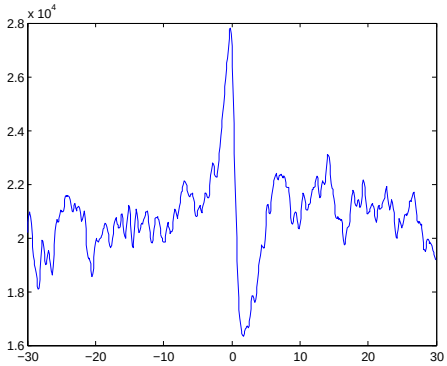

Figure 4: Cumulative number of head nods (Y axis) in function of the time alignment with all contextual events (X axis): spoken words, prosodic and timing. We can observe a relationship between contextual events and head nods between -5 and 5 seconds.

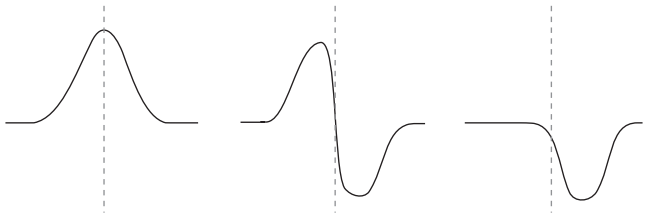

Figure 5: Schematic representation of the three patterns observed when analyzing co-occurrence of head nods and contextual events: (left) ignition pattern, (middle) transition pattern and (right) negative pattern.

3.2 Patterns in Co-occurrence Graphs

By observing the co-occurrence graphs of Figure 3, three patterns appear: ignition, transition and negation. These patterns are illustrated in Figure 5.

- **Ignition pattern** The first pattern is the ignition pattern (left) where a contextual event positively influence visual gesture. This type of relationship means that a visual gesture is more likely to happen around the contextual event. This is true for the period which represents the end of a sentence. This is also true for positive feed such as the word "yeah".
- **Transition pattern** The second pattern is the transition pattern (middle) where a contextual event represents a mid-point between two phrases. This type of relations will bring a high likelihood around or before the event but right after the event this likelihood will be lower. Two good examples of this type of events are the comma and the word "and". These

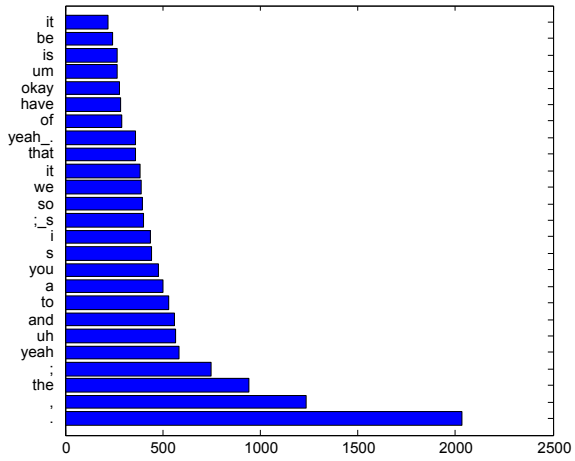

Figure 6: Top 25 contextual features. Horizontal axis: maximum number of time a head nod happened in a window or +/-5 seconds around the contextual feature.

events will usually occur in the middle of a sentence, between two constituents.

- **Negative pattern** The last pattern is the "negative" pattern (right) where a contextual event negatively influence a visual gesture. This type of relations means that a visual gesture is unlikely to happen after this event. The words "the" and "to" are two good examples of this type of patterns. These words do not bring visual feedback and usually following one of these words will be a large number of other spoken words.

The analysis of the co-occurrence graphs shown in Figure 3 confirm our intuition that the context is related to visual feedback (e.g., head nods). Also, these co-occurrence graphs contains patterns that can potentially help to recognize when a specific gesture is more likely.

3.3 Co-occurrence Ranking of Contextual Features

A good contextual feature is an event (1) that happens on a regular basis so that there is a good chance to see this same event in a new meeting, and (2) that is related to visual feedback. One criterium that includes both advantages is the maximum number of co-occurrence between the contextual event and the visual gesture. This criteria is equal to the maximum peak of each co-occurrence graphs.

Figure 6 shows the top 25 contextual features. The top feature is the period, which usually represent the end of a sentence. This goes with our intuition that

people usually do grounding gesture at the end of a sentence. Also the second feature is the comma which represents a pause in a sentence. Pauses are also good timing for grounding gesture. The other top contextual features are more interesting since they are lexical features and bring interesting questions as why they are related with visual gestures.

4 Leaning Relevant Context

4.1 Sequential Probabilistic Model

A sequential probabilistic model takes as input a sequence of observation features (e.g., the speaker features) and returns a sequence of probabilities (i.e., probability of listener backchannel). Two of the most popular sequential models are Hidden Markov Model (HMM) [13] and Conditional Random Field (CRF) [10]. One of the main difference between these two models is that CRF is discriminative (i.e., tries to find the best way to differentiate cases where the listener gives backchannel to cases where it does not) while HMM is generative (i.e., tries to find the best way to generalize the samples from the cases where the listener gives backchannel without looking at the cases where the listener did not give backchannel). Our prediction model is designed to work with both types of sequential probabilistic models.

Sequential probabilistic models such as HMM and CRF have some constraints that need to be understood and addressed before using them:

- **Limited learning** The more informative your features are, the better your sequential model will perform. If the input features are too noisy (e.g., direct signal from microphone), it will make it harder for the HMM or CRF to learn the important part of the signal. Also, because of the pre-processing your input features to highlight their influences on your label (e.g., listener backchannel) you improve your chance of success.

- **Over-fitting** The more complex your model is, the more training data it needs. Every input feature that you add increases its complexity and at the same time its need for a larger training set. Since we usually have a limited set of training sequences, it is important to keep the number of input features low.

In our prediction model we directly addressed these issues by focusing on the feature representation and feature selection problems:

- **Encoding dictionary** To address the limited learning constraint of sequential models, we suggest to use more than binary encoding to represent input features. Our encoding dictionary contains a series of encoding templates that were designed to model different relationship between a speaker feature (e.g., a speaker in not currently speaking) and listener backchannel. The encoding dictionary and its usage are described in Section 4.2.

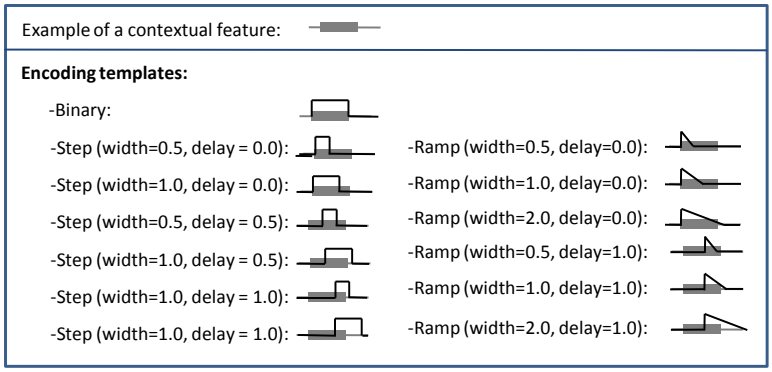

Figure 7: **Encoding dictionary.** This figure shows the different encoding templates used by our context-based approach. Each encoding template was selected to express a different relationship between contextual features and visual feedback. This encoding dictionary gives a more powerful set of input features to the sequential probabilistic model and improves the performance of our context-based recognizer.

- **Automatic feature and encoding selection** Because of the over-fitting problem happening when too many uncorrelated features (i.e., features that do not influence listener backchannel) are used, we suggest two techniques for automatic feature and encoding selection based on co-occurence statistics and performances evaluation on a validation dataset. Our feature selection algorithms are described in Section 4.3.

The following two sections describe our encoding dictionary and feature selection algorithm.

4.2 Encoding Dictionary

The goal of the encoding dictionary is to propose a series of encoding templates that capture the coarse relationship between speaker features and listener backchannel. These encoding templates will help to represent long-range dependencies (when the influence of an input feature decay slowly, possibly with a delay) that are otherwise hard to learn using a sequential probabilistic model. An example of a long-range dependency will be the effect of low-pitch regions on backchannel feedback with an average delay of 0.7 seconds (observed by Ward and Tsukahara citeward00). In our framework, the prediction model will pick an encoding template with a 0.5 seconds delay and the exact alignment will be learned by the sequential probabilistic model which will also take into account the influence of other input features.

The Figure 7 shows the 13 encoding templates used in our experiments. These encoding templates were selected to represent a wide range of ways that

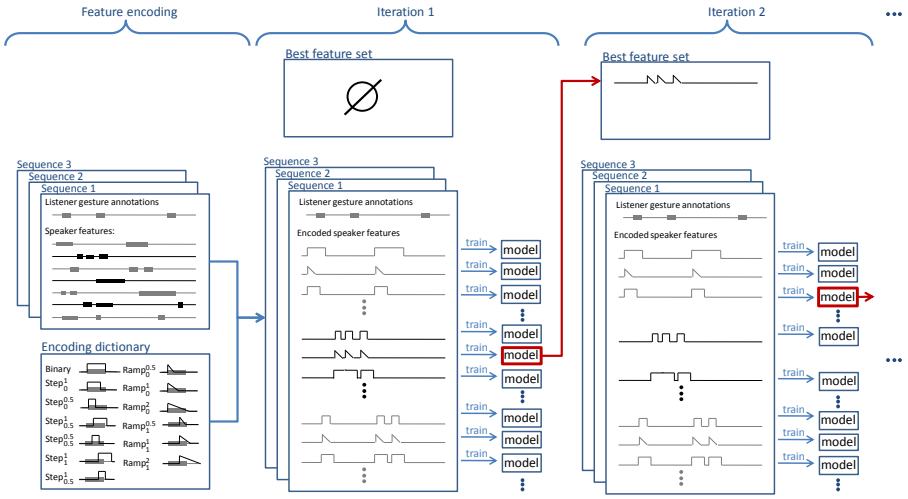

Figure 8: Joint feature selection. This figure illustrates the feature encoding process using our encoding dictionary as well as two iterations of our joint feature selection algorithm. The goal of joint selection is to find a subset of features that best complement each other for recognition of listener visual feedback.

a speaker feature can influence the listener backchannel. These encoding templates were also selected because they can easily be implemented in real-time since the only needed information is the start time of the speaker feature. Only the binary feature also uses the end time. In all cases, no knowledge of the future is needed.

The three main types of encoding templates are:

- **Binary encoding** This encoding is designed for speaker features which influence on listener backchannel is constraint to the duration of the speaker feature.
- **Step function** This encoding is a generalization of binary encoding by adding two parameters: width of the encoded feature and delay between the start of the feature and its encoded version. This encoding is useful if the feature influence on backchannel is constant but with a certain delay and duration.
- **Ramp function** This encoding linearly decreases for a set period of time (i.e., width parameter). This encoding is useful if the feature influence on backchannel is changing over time.

It is important to note that a feature can have an *individual* influence on backchannel and/or a *joint* influence. An *individual* influence means the input feature directly influences listener backchannel. For example, a long pause can

by itself trigger backchannel feedback from the listener. A *joint* influence means that more than one feature is involved in triggering the feedback. For example, saying the word "and" followed by a look back at the listener can trigger listener feedback. This also means that a feature may need to be encoded more than one way since it may have a *individual* influence as well as one or more *joint* influences.

One way to use the encoding dictionary with a small set of features is to encode each input feature with each encoding template. We tested this approach in our experiment with a set of 12 features (see Section 5) but because of the problem of over-fitting, a better approach is to select the optimal subset of input features and encoding templates. The following section describes our feature selection algorithm.

4.3 Joint Feature Selection

Given the subset of features that performed best when trained individually, we now build the complete set of feature hypothesis to be used by the joint feature selection process. This set represents each feature encoded with all possible encoding templates from our dictionary. The goal of joint selection is to find a subset of features that best complements each other for prediction of backchannel. Figure 8 shows the first two iterations of our algorithm.

The algorithm starts with the complete set of feature hypothesis and an empty set of *best* features. At each iteration, the best feature hypothesis is selected and added to the best feature set. For each feature hypothesis, a sequential model is trained and evaluated using the feature hypothesis and all features previously selected in the best feature set. While the first iteration of this process is really similar to the individual selection, every iteration afterward will select a feature that best complement the current best features set. Note that during the joint selection process, the same feature can be selected more than once with different encodings. The procedure stops when the performance starts decreasing.

5 Examples of Affective Behavior Understanding

In this section, we show the importance of context in affective behavior understanding on two different domains: interaction with a robot and human dyadic interactions. Our hypothesis is that the inclusion of contextual information within the affective gesture recognizer will increase the number of recognized gestures while reducing the number of false detections. In both domains we compares two different configurations: (1) using the vision-only approach, and (2) combining the contextual information with the results of the visual approach (multi-modal integration).

Figure 9: Mel, the interactive robot, can present the iGlassware demo (table and copper cup on its right) or talk about its own dialog and sensorimotor abilities.

5.1 Robot Demonstration

The following experiment demonstrates how contextual features inferred from an agent's spoken dialogue can improve head nod and head shake recognition. The experiment compares the performance of the vision-only recognizer with the context-only prediction and with multi-modal integration.

Human participants were video recorded while interacting with the robot (see Figure 10). The vision-based head tracking and head gesture recognition was run online (∼18Hz). The robot's conversational model, based on COLLAGEN [14], determines the next activity on the agenda using a predefined set of engagement rules, originally based on human–human interaction [16]. Each interaction lasted between 2 and 5 minutes. During each interaction, we also recorded the results of the vision-based head gesture recognizer as well as the contextual cues (spoken utterances with start time and duration) from the dialogue manager. These contextual cues were later automatically processed to create the contextual features (see Section 2.1) necessary for the contextual predictor. For ground truth, we hand labeled each video sequence to determine exactly when the participant nodded or shook his/her head. A total of 274 head nods and 14 head shakes were naturally performed by the participants while interacting with the robot.

Figure 9 shows head nod detection results for all 9 subjects used during testing. The ROC curves present the detection performance each recognition algorithm when varying the detection threshold. The area under the curve for each techniques are 0.9482 for the vision only, 0.7691 for the predictor and 0.9678 for the integrator. For head shakes, theareas under the curve for each

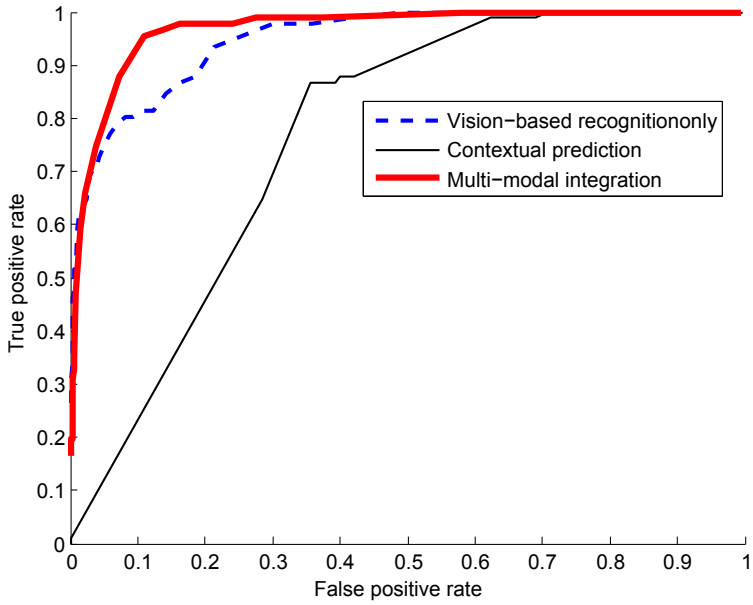

Figure 10: Results for robot demonstration: head nod recognition curves when varying the detection threshold. For a fixed false positive rate of 0.0409 (operating point), the context-based approach improves head nod recognition from 72.5% (vision only) to 90.4%.

techniques are 0.9780 for the vision only, 0.4961 for the predictor and 0.9872 for the integrator.

Table 1 summarizes the results for head nods and head shakes understanding by computing the true positive rates for the fixed negative rate of 0.1. Using a standard analysis of variance (ANOVA) on all the subjects, results on the head nod detection task showed a significant difference among the means of the 3 methods of detection: $F(1,8) = 62.40$, $p < 0.001$, $d = 0.97$. Pairwise comparisons show a significant difference between all pairs, with $p < 0.001$, $p = 0.0015$, and $p < 0.001$ for vision-predictor, vision-integrator, and predictor-integrator respectively. A larger number of samples would be necessary to see the same significance in head shakes.

	Vision	Predictor	Integrator
Head nods	81%	23%	**93%**
Head shakes	83%	10%	**98%**

Table 1: True detection rates for a fix false positive rate of 0.1.

Figure 11: Setup for Human dyadic storytelling. This study of face-to-face narrative discourse (i.e., quasi-monologic storytelling) included 76 subjects. The speaker was instructed to retell the stories portrayed in two video clips to the listener.

5.2 Human Dyadic Story Telling

Data is drawn from a study of face-to-face narrative discourse ('quasi-monologic' storytelling). 76 subjects from the general Los Angeles area participated in this study. Participants in groups of two entered the laboratory and were told they were participating in a study to evaluate a communicative technology. Subjects were randomly assigned the role of speaker and listener. The speaker viewed a short segment of a video clip taken from the Edge Training Systems, Inc. Sexual Harassment Awareness video. After the speaker finished viewing the video, the listener was led back into the computer room, where the speaker was instructed to retell the stories portrayed in the clips to the listener. The listener was asked to not talk during the story retelling. Elicited stories were approximately two minutes in length on average. Participants were debriefed individually and dismissed.

We collected synchronized multimodal data from each participant including voice and upper-body movements. Both the speaker and listener wore a lightweight headset with microphone. Three Panasonic PV-GS180 camcorders were used to videotape the experiment: one was placed in front the speaker, one in front of the listener, and one was attached to the ceiling to record both speaker and listener

Our first experiment compared the performance of our context-based recognition framework with a vision-only recognizer. Figure 12 shows the ROC curve for both approaches. The ROC curves present the detection performance for both recognition algorithms when varying the detection threshold. The two quantitative methods used to evaluate ROC curves are area under the curve and equal error rate. Table 2 shows the quantitative evaluation using both error criteria. The use of context improves recognition from 74.9% to 83.2%.

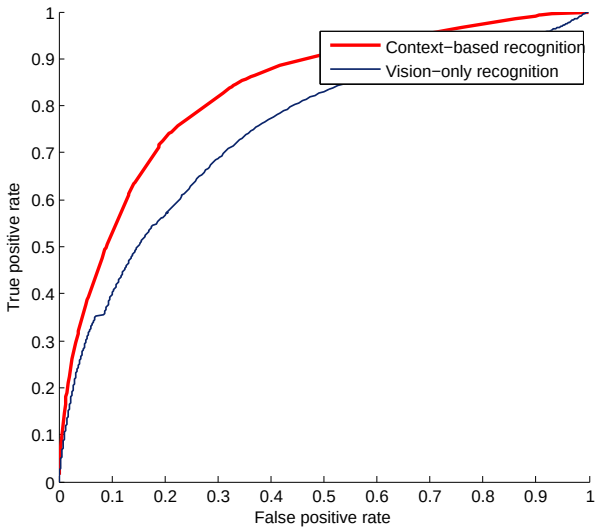

Figure 12: Results for human dyadic storytelling: ROC curves of head nod recognition comparing our context-based approach to a vision-only approach.

Recognizer	Area	EER
Context-based	83.2%	76.5%
Vision-only	74.9%	69.4%

Table 2: Quantitative comparison between our context-based approach and a vision-only approach (same as Figure 12). The table shows both the area under the curve and the equal error rate (EER).

Pairwise two-tailed t-test comparison show a significant difference for both error criteria, with $p = 0.021$ and $p = 0.012$ for the area under the curve and the equal error rate respectively.

As described in Section 4, our context-based recognition framework uses two types of feature selections: individual feature selection and joint feature selection (see Section 4.3 for details). It is very interesting to look at the features and encoding selected after both processes:

- *Vowel volume* using ramp encoding with a width 0.5 second and a delay of 0.5 seconds
- *Speaker looking at the listener* using a binary
- *Pause* using step encoding with a width 2.0 second and no delay
- *Low pitch* using ramp encoding with a width 0.5 second and no delay

Feature selection	Area	EER
Joint + Individual	83.2%	76.5%
Individual only	79.1%	72.0%

Table 3: Quantitative evaluation showing the gain in performance when using both individual and joint feature selection.

Feature encoding	Area	EER
Dictionary	79.1%	72.0%
Binary	76.1%	69.9%

Table 4: Quantitative evaluation showing the gain in performance when using the encoding dictionary for feature representation.

These are the four features-encoding pairs selected after the joint feature selection process which stopped when validation performance started decreasing. We can see that only one feature was selected with binary encoding, suggesting that the use of the encoding dictionary was important. The first selected feature *Vowel volume* used an encoding with a ramp and a delay of 0.5 seconds, meaning that its influence on head nods is asynchronous and decreases over time. The second selected feature is related to the eye gaze of the speaker, confirming the importance of our multimodal context. The third and fourth features have also been reported by Ward and Tsukahara [19] as good predictive features for backchannel feedback. No lexical feature was selected by the *joint* selection algorithm. This result means that visual gesture recognition can be improved using only prosodic cues, pauses and speaker visual display.

The second and third experiments were designed to understand the influence of feature selection and encoding dictionary on the context-based recognition framework. Table 3 compares the recognition performance when using or not using the joint feature selection after the individual feature selection. Table 4 compares the recognition performance when using the complete encoding dictionary to using only binary encoding. This last comparison was done after the individual feature selection.

We can see from both Table 3 and 4 that the gain performance of our context-based recognition algorithm is directly related to the joint feature selection and the encoding dictionary. By using the encoding dictionary instead of the usual binary encoding, the performance improves from 76.1% to 79.1%. And by using the joint feature selection, the performance improves again from 79.1% to 83.2%.

Our experiments show that by using joint feature selection and an encoding dictionary, contextual information from other participant significantly improve the performance of vision-based gesture recognition.

6 Conclusion

Our results show that contextual information from other human participants can improve visual gesture recognition. We presented a context-based recognition framework that represents contextual features based on an encoding dictionary and automatically selects the optimal features based on *individual* and *joint* influence. By using simple prosodic, pauses and visual display contextual features available in real-time, we were able to improve the performance of the vision-only head gesture recognizer from 74.9% to 83.4%. An important contribution of this chapter was our contextual representation based on co-occurrence graphs which models co-occurrence between contextual cues such as spoken words and pauses, and visual head gestures. By analyzing these co-occurrence patterns we automatically selected relevant contextual features and predicted when visual gestures was most likely. Recognizing these visual gestures is important for understanding the full meaning of a meeting or conversation, and can be particularly crucial for identifying more subtle factors such as the effectiveness of communication or diagnosis social disorders.

References

[1] J. Allwood, J. Nivre, and E. Ahlsn. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, pages 1–26, 1992.

[2] Janet B. Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[3] Judee K. Burgoon, Lesa A. Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press, Cambridge, 1995.

[4] Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 1999.

[5] N. Cathcart, Jean Carletta, and Ewan Klein. A shallow model of backchannel continuers in spoken dialogue. In *European ACL*, pages 51–58, 2003.

[6] H. H. Clark. *Using Language.* Cambridge University Press, 1996.

[7] AMI consortium. *AMI meeting corpus.* http://corpus.amiproject.org/.

[8] A. Dittmann and L. Llewellyn. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9:79–84, 1968.

[9] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[10] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

[11] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labelling sequence data. In *ICML*, 2001.

[12] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence*, 171(8–9):568–585, June 2007.

[13] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. A spoken dialog system for chat-like conversations considering response timing. *LNCS*, 4629:599–606, 2007.

[14] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[15] Rich, Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human–computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*, 22(4):15–25, 2001.

[16] A.A. Rizzo, D. Klimchuk, R. Mitura, T. Bowerly, J.G. Buckwalter, and T. Parsons. A virtual reality scenario for all seasons: The virtual classroom. *CNS Spectrums*, 11(1):35–44, 2006.

[17] C. Sidner, C. Lee, C.D.Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164, August 2005.

[18] Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293, 1990.

[19] Larissa Z. Tiedens and Alison R. Fragale. Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology*, 84(3):558–568, 2003.

[20] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

[21] V. H Yngve. On getting a word in edgewise. In *Sixth regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.